Scalable Resume Analytics with the Databricks Lakehouse

A Big Data Pipeline for Talent Intelligence & Market Insights

FAHIM THANZEEL
24MBMB16

The Solution: Medallion Architecture

BRONZE Layer

Ingests all raw, unstructured resume text files (`.txt`) directly from the source. This layer provides a reliable, versioned "Source of Truth" saved as a Delta Lake table (`bronze_resumes`).

SILVER Layer

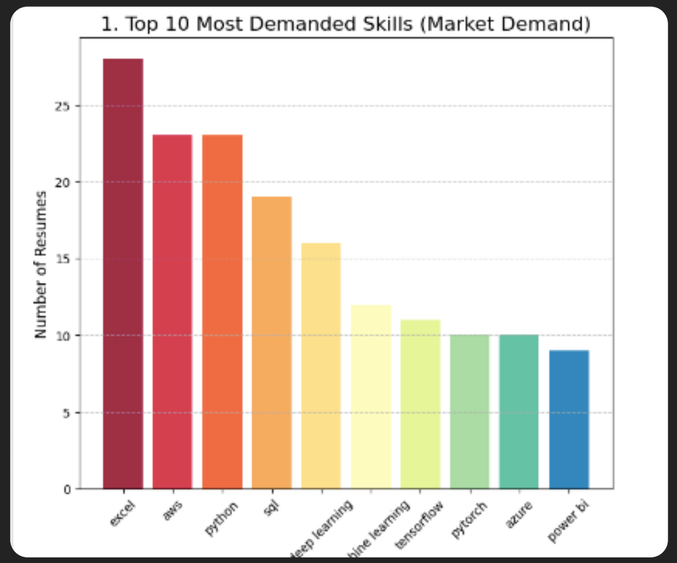
Cleanses, enriches, and transforms the data. A PySpark UDF (User Defined Function) with Regular Expressions is applied to extract structured features like `extracted_skills`, creating the analysis-ready `silver_features` table.

GOLD Layer

The final business-facing layer. It aggregates Silver data to create consumable reports, ML model outputs, and visualizations for the dashboard. All five business use cases are served from this layer.

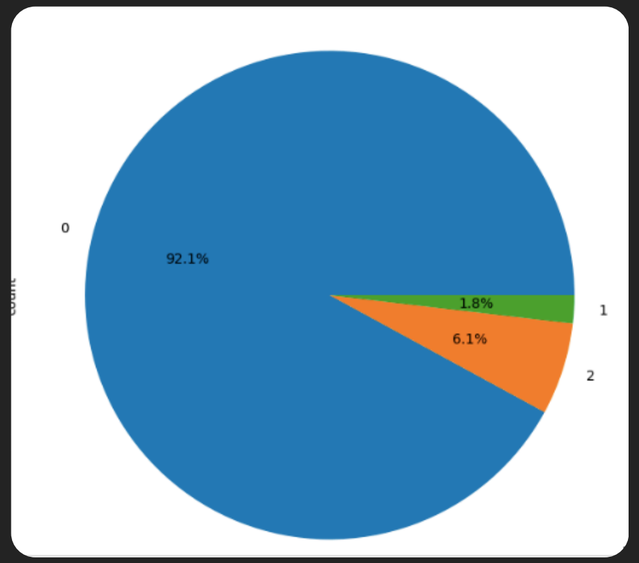
Core Analytics: Market & Candidate Insights

1. Skill Demand Analytics



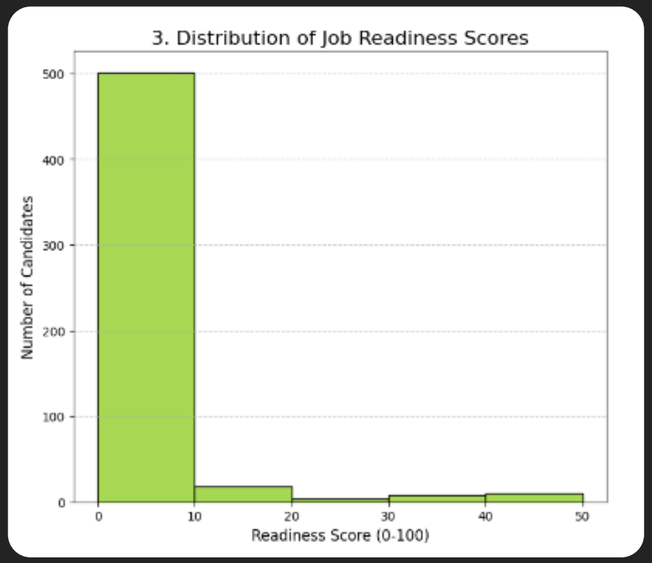
Identifies the most frequent technical skills found in the applicant pool, guiding curriculum and hiring focus.

2. Talent Pool Composition



Uses K-Means clustering (PySpark MLlib) to segment candidates into distinct, human-readable talent groups.

3. Job Readiness Scoring

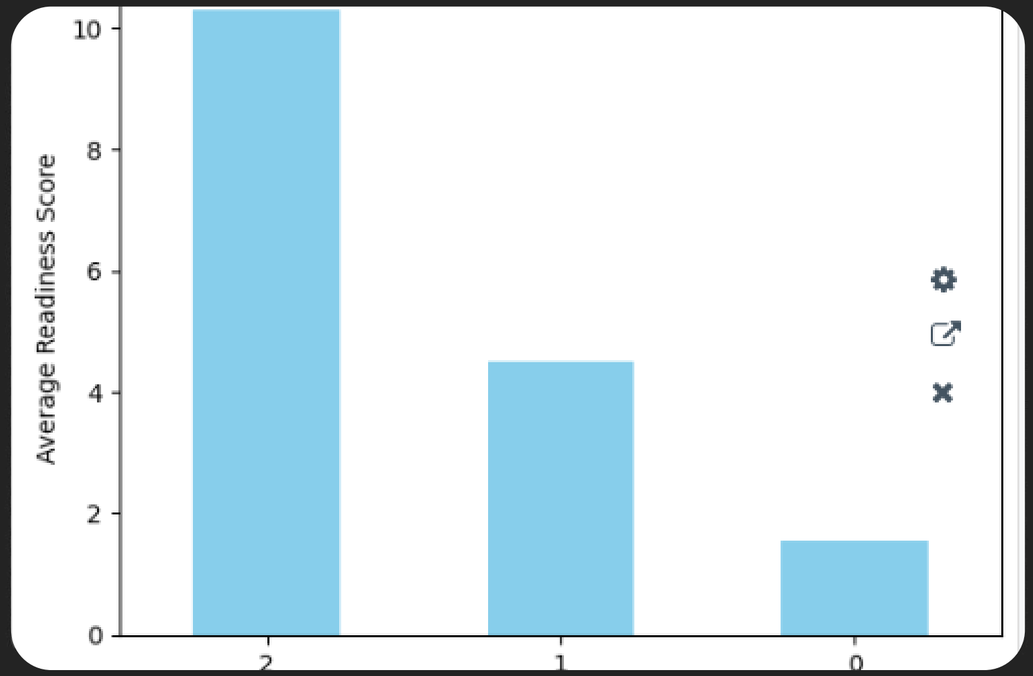


Applies a rule-based ATS score to every resume, providing a 0-100 metric for ranking and shortlisting candidates.

Advanced Insights: Strategy & Skill Gaps

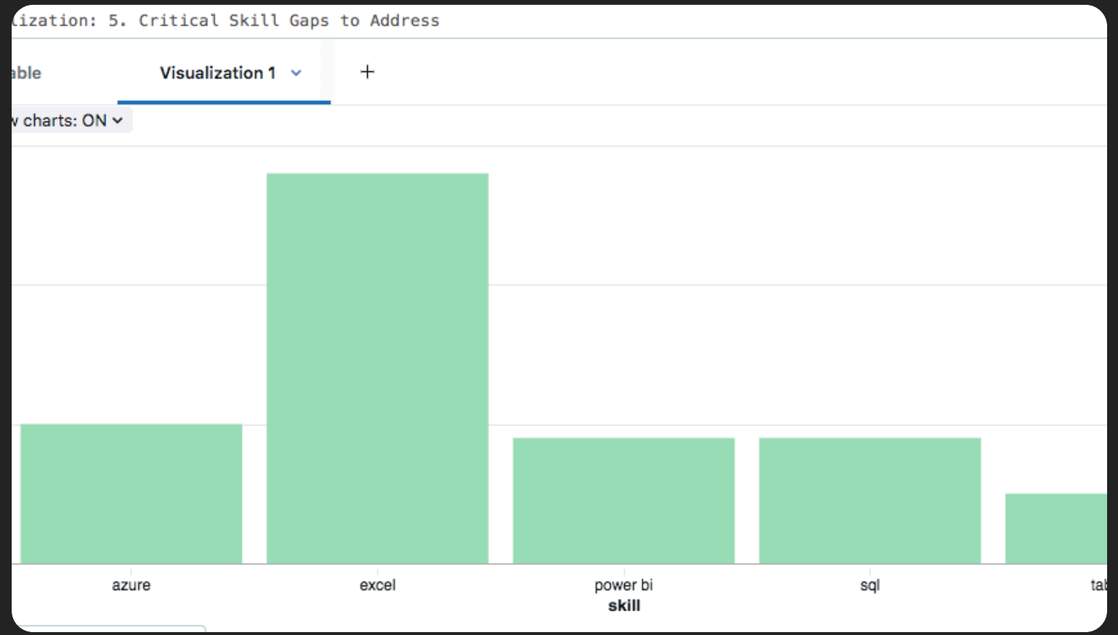
4. Market Intelligence Report

This report combines Use Cases 2 and 3 to find the average 'Readiness Score' for each 'Talent Segment'.



5. Skill Gap Analysis

This analysis identifies the critical skills that the high-performing segments have, but the low-performing segments lack.



Conclusion & Future Scope

- Successfully built a scalable **end-to-end Big Data pipeline** using the Medallion Architecture on Databricks.
- Automated the entire workflow (Bronze → Silver → Gold) using a **Master Pipeline Runner** (`dbutils.notebook.run`) for one-click refresh.
- Generated **five distinct, actionable business insights** from raw text, including ML-driven segmentation and prescriptive skill gap analysis.

thank you